Proposal to Promote Improved Bacterial Genomic Annotation.

This is a proposal for a collaborative effort between NCBI, ASM, and sequencing centers to improve the annotation of bacterial genome submissions. This includes the proper use of locus_tags, gene symbols and protein names so that a continuity exists between the published literature and what is found within various databases containing genomic data.

I. Proper use of locus_tags in bacterial genomes.

1. Current state.

Locus_tags were intended to be used for the purpose of tracking genes in a single submission. An example would be the Blattner numbers (b numbers) used in Escherichia coli K-12 MG1655 (GenBank Accession U00096; RefSeq Accession NC_000913). The idea was that locus_tags would be unique in a single genomic submission. As the number of bacterial genome submissions has increased, it has become apparent that there are increasing problems with regard to the use of locus_tags:

1A. Locus_tags continue to be unique within a given genome submission.

1B. Locus_tags are not unique across genomes. There are multiple instances of a given locus_tag referring to two different genes in two organisms. (see Appendix)

1C. Locus_tags have become surrogate gene symbols. They are used to identify and discuss genes when a gene symbol has not yet been officially established. In both the literature and in submissions to the nucleotide databases existing locus_tags are used to denote similarity.

As the number of genome submissions will increase in the future, it is likely that these problems will increase both in scope and number.

2. Proposal to improve usage.

NCBI and ASM propose that a database be set up to both register and check for the use of locus_tags (especially locus_tag prefixes) so that these types of problems are prevented in the future. This database will be publicly available and searchable. Searches can be made for publicly available locus_tags (or prefixes) or for reserved locus_tags (or prefixes). Searches that find locus_tags that are publicly available will provide information about the published genome while those for reserved locus_tags (or prefixes) will provide information that the locus_tag is reserved without reference to the organism or sequencing center. Sequencing centers or submitters may also register their locus_tags ahead of time by registering either a prefix or prefix and number range prior to submitting genome information. The intent is that this will prevent the types of locus_tag

problems that occur now especially when multiple genome submissions occur simultaneously and that contain duplicate locus_tags.

3. Proper usage of locus_tags.

Locus_tags are also problematic for another reason. Depending on the framework for their construction, they may potentially cause confusion with other alphanumeric identifiers used at GenBank/DDBJ/EMBL, namely Accession Numbers that consist of 3 letters+5 numbers or 2 letters+6 numbers (see below). We wish to avoid this sort of confusion and will work to prevent locus_tags with the above pattern from appearing in genome submissions. Coupled with the above points in #1, NCBI proposes the following standards be adopted.

3A. Locus_tags should be unique within a given whole genome submission (a whole genome submission are for complete submissions, as for WGS, submissions). Note that locus_tag usage on single genes or small genomic fragments will remain uncontrolled.

3B. Locus_tags should be unique across all whole genomic submissions as per 3A.

3C. Locus_tags should use a symbol to separate the prefix from the numeral. We propose using underscore (_) for this as it is easily searchable. This will also prevent confusion with Accession Numbers. (ex. ABC_00001).

3D. Locus_tag prefixes should be equal to or greater than 3 alphanumeric characters. They should not start with a numeral, but numerals can be in the 2nd position or later in the string. (ex. A1C_00001)

3E. The qualifier old_locus_tag should be used for tracking purposes if a locus_tag is changed on a publicly available genome.

3F. If a genome submission is updated, identical genes should carry the same locus_tags. See below for discussions on merged/split genes.

3F. If a genome submission is updated, new genes should use the same prefix with a new numerical identifier. The new locus_tag could be incremented from the last locus_tag in the original submission. Use of decimal integers will be discouraged as it mimics version numbers. For genomes that may undergo multiple revisions, typically eukaryotic, it might be useful to promote the use of large integer gaps between locus_tags to easily allow the addition of new genes in subsequent revisions.

USE ONE OF THE FOLLOWING:

3F1. Incremental locus_tags

Original submissions	revised submission
ABC_0022	ABC_0022
	ABC_4568 (new gene)
ABC_0023	ABC_0023

OR

3F2. Gaps in original locus_tags

Original submissions	revised submission
ABC_0020	ABC_0020
	ABC_0021 (new gene)
ABC_0030	ABC_0030

BUT NOT

3F3. Decimal integers

Example from Cryptococcus genome:

ABC_0020	ABC_0020
	ABC_0020.1 (new gene)
ABC_0030	ABC_0030

3G. Genome centers SHOULD NOT use the same locus_tag prefixes for multiple submissions of different organisms. Using the genome center prefix may be easier for the genome center but is harder on the users as they expect a specific prefix to be associated with a specific organism.

3H. These recommendations should apply to both prokaryotic and eukaryotic genome submissions. Those users wishing to encode information on chromosome number, etc., can apply that information to the locus_tag after the prefix. ex. ABC_I00001 for gene 1, chromosome 1, and ABC_II00001 for gene 1, chromosome 2. Similarly this can apply to prokaryotic genomes when there are multiple chromosomes or a chromosome and plasmids, and to marking genes as tRNAs or rRNAs. However, experience shows that it is generally not wise to encode functional information, including chromosome number, into locus_tags as this information may change at a later date, resulting in either degradation of the meaning in the locus_tag, or the necessiry of a change to the locus_tag even though the gene remains the same.

3I. NCBI will provide a searchable database for existing locus_tags. Searches can be done with taxname, taxid, Accession, gi, locus_tag prefix, and locus_tag. In prototype stage.

3J. NCBI will provide a registry so that sequencing centers can register their locus_tag prefix. Not yet implemented.

3K. NCBI will then implement searching of confidential locus_tags after 3J is finished. If a user initiates a search for a confidential locus_tag prefix, they will be informed that the locus_tag prefix is currently registered, but no other information will be returned to the user. Not yet implemented.

4. Items for discussion - locus_tags.

With regard to the usage of locus_tags and the locus_tag database that NCBI and ASM propose, there are several areas which need to be thought about prior to discussion.

4A. What should get a locus_tag? All coding regions and RNAs?

4B. Genome updates and tracking of locus_tags.

-what happens when a genome gets updated and genes are merged, or split

4C. New genome submissions

-new genome submissions should get new and unique locus_tags

4D. How to deal with problems and conflicts

-enforcement

4E. Registering locus_tag prefixes and number ranges.

4F. Submission of proposal to DDBJ/EMBL prior to the International Nucleotide Sequence Database collaboration meeting in May, 2005.

4G. Submission of abstract to the International Conference on Microbial Genomes (abstract due Feb. 14th - meeting April 13-16) to propose this to the major sequencing centers (http://www.tigr.org/conf/mg2005/)

4H. Use of locus_tags as surrogate gene symbols. If these numbers may be used as surrogate temporary gene names, then a means for distinguishing mutant alleles is needed. This could be something as simple as writing (in a genotype) the locus_tag in parenthesis followed by an allele number; e.g., (ABC_0001)26 or for a gene fusion, *lacZ::(ABC_0001)2*; the locus_tag might be in italics or not as a matter of convention.. Also see the next point. The locus_tag convention should be compatible with such usage. Similar abbreviations could be created for presumptive RNA, such as *prn,* or *psi* (presumptive site). Having a way to systematically name mutants (or natural varients) is very important to the concept underlying bacterial genes names, as well as to the reality that people involved in making varients of any patentable protein (e.g., toxin genes) really want to be able to have a firm way of identifying varients.

4I. Enforcement of the usage of locus_tags in ASM journals.

4J. Possible collision of locus_tags with existing gene symbols. This is not so much of a problem with prokaryotes, but it is with eukaryotes.

4K. What to do with existing locus_tags? Are they to be moved to the qualifier 'old locus_tag' and new ones fitting the proposal above created in their place?

5. Collaboration with ASM.

NCBI wishes to collaborate with ASM on setting up this database in order to foster communication between both NCBI and ASM, ASM and the major sequencing centers, and NCBI and the major sequencing centers. The intent is that initial collaboration on locus_tags between all 3 will foster communication that will improve the annotation of bacterial genomes in other ways, namely for systematic gene symbols and protein names.

6. Future directions.

In our efforts to improve the annotation of bacterial genomes there are areas in which we hope to make progress. As an example of the types of improvements NCBI wishes to discuss, the use of various flavors of "hypothetical protein" as a protein name such as "predicted protein", "hypothetical orf", etc. NCBI proposes to reduce the range of names used in bacterial annotation and simply use "hypothetical protein" for all proteins that do not show similarity to an experimentally characterized protein. NCBI is open to discussion on this and other areas in which the annotation of genomes can be improved upon.

Appendix.

Multiple Locus_tag Prefixes

The following organisms have genome-wide duplicate locus_tag prefixes:

Organism	Acc	prefix
Staphylococcus aureus subsp. aureus Mu50	BA000017	SAV
Streptomyces avermitilis MA-4680	BA000030	SAV
Borrelia burgdorferi B31	AE000783	BB
Bordetella bronchiseptica RB50	BX470250	BB
Bacteroides fragilis NCTC 9343	CR626927	BF
Bacteroides fragilis YCH46	AP006841	BF
Shigella flexneri 2a str. 301 plasmid pCP301	AF386526	CP
Chlamydophila pneumoniae AR39	AE002161	CP

Multiple Locus_tags

Other sets of locus_tags are used in multiple instances. Typically these are for structural RNAs.

ex:

tRNA-Val-6
rRNA-16SrRNA_1
RNA_20

tRNA-Val-6

Bacillus anthracis str. Ames	AE016879

Bacillus anthracis str. 'Ames Ancestor'	AE017334
Bacillus cereus ZK	CP000001
Bacillus anthracis str. Sterne	AE017225
Vibrio vulnificus CMCP6	AE016796
Bacillus thuringiensis serovar konkukian str. 97-27	AE017355

Feature Table

International Nucleotide Sequence Database Collaboration Feature Table Definition (DDBJ/EMBL/GenBank).

http://www.ncbi.nlm.nih.gov/projects/collab/FT/index.html
(updated Oct. 2004)

Current definitions:

Feature Key **gene**

Definition region of biological interest identified as a gene and for which a name has been assigned;

Optional qualifiers
/allele="text"
/citation=[number]
/db_xref="<database>:<identifier>"
/evidence=<evidence_value>
/function="text"
/gene="text"
/label=feature_label
/locus_tag="text" (single token)
/map="text"
/note="text"
/old_locus_tag="text" (single token)
/operon="text"
/product="text"
/pseudo
/phenotype="text"
/standard_name="text"
/usedin=accnum:feature_label

Comment the gene feature describes the interval of DNA that corresponds to a genetic trait or phenotype; the feature is, by definition, not strictly bound to it's positions at the ends; it is meant to represent a region where the gene is located.

Qualifier **/gene=**
Definition symbol of the gene corresponding to a sequence region
Value format "text"
Example /gene="ilvE"

Qualifier **/locus_tag**

```
Definition      feature tag assigned for tracking purposes
Value Format    "text"(single token)
                but not "<1-5 letters><5-9 digit integer>[.<integer>]"
Example         /locus_tag="RSc0382"
                /locus_tag="YPO0002"
```
Comment /locus_tag can be used with any feature where /gene is valid;
identical /locus_tag values may be used within an entry/record, but only if the identical /locus_tag values are associated with the same gene; in all other circumstances the /locus_tag value must be unique within that entry/record. Multiple /locus_tag values are not allowed within one feature for entries created after 15-OCT-2004.
If a /locus_tag needs to be re-assigned the /old_locus_tag qualifier should be used to store the old value. Existing records where multiple /locus_tag qualifiers are present will be retrofitted by January 2005.
The /locus_tag value should not be in a format which resembles INSD accession numbers, accession.version, or /proteid_id identifiers.